Assignment 1

CSL7620: Machine Learning

AY 2023-24, Semester – I

Due on: 27/08/2023 11:59 pm

M.M: 80

General Instructions:

1. Clearly mention the assumptions you have made, if any.

2. Clearly report any resources you have used while attempting the assignment.

3. Any submission received in another format or after the deadline will not be evaluated.

4. Make sure to add references to the resources that you have used while attempting the assignment.

5. Plagiarism of any kind will not be tolerated and will result in zero marks.

6. Select your dataset correctly. If found otherwise, your assignment will not be evaluated.

Submission Guidelines:

1. Prepare separate Python code files for each task and name them YourRollNo_task1.py, YourRollNo_task2.py, and so on.

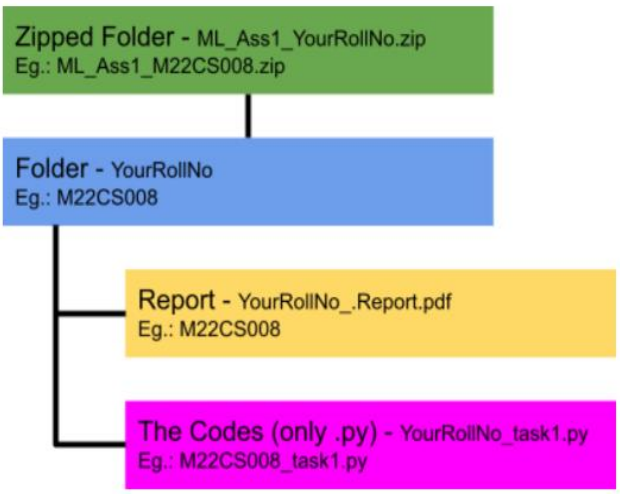
2. Also, provide your colab file link in the report. Make sure that the file is sharable.

3. Put both the codes and a report in a folder named <YourRollNo>, create a zip folder named ML_Ass1_YourRollNo.zip, and upload to google-classroom. See attached image to get better clarity.

4. Do not download the .ipynb file, rename it as .py, and upload it. .ipynb files are not exactly in a readable form, so uploading it will only result in you receiving 0 marks for the same. You have an option to download .py file in google colab. Use it to get the .py format.

5. Submit a single report depicting all tasks' methods, results, and observations. There is no need to add theory behind the concepts. Preparing a report is mandatory; failing it will lead to non-evaluation of the assignment.

6. Do not copy-paste code or screenshots, etc. in the report. The report should look like a technical document, containing plots, tables, etc. whenever necessary.



Dataset Link:
https://drive.google.com/file/d/1Ik1JdRC9eWljtclz_WkYmDEJA7cvcxJY/view?usp=sharing

Question 1:

The Student Performance Dataset is a dataset designed to examine the factors influencing academic student performance. Based on the data available we want to find out the student's performance given some attribute values. Download the give dataset and perform the followings – [60 marks]

1. Identify the dependent and independent variables. **2 marks.**
2. Read the dataset and do exploratory data analysis. (Data Preprocessing and meaningful plots) **8 marks.**
3. Split the data set in train and test (80:20) ratio. **2 marks.**
4. Write a python code for Linear Regression (from **scratch**) and train the model with training data. (Only **numpy** and **pandas** should be used) **30 marks**
5. Plot the loss vs epoch curve. **3 marks**
6. **Give a student's data – [**Hours of study = 7, Previous score = 95, Extracurricular Activities = Yes, Duration of Sleep = 7, Sample Question Papers Practiced = 6**]** then What will be his/her performance based on your trained model. **2 marks**
7. **Evaluate the model's performance based on any two-performance metrics (at least 2) from below on the test set –** a.) MSE error b.) R2 Score c.) Adjusted R2 score (Only numpy and pandas is allowed) **5 marks**

Question 2:

Take the same dataset and do it using regression library available in python. Analyse and compare the result of your model (from scratch) vs library created model. Can you improve your model's performance and how? [**20 marks]**

Note: Include a single report for above questions. The report should be crisp and compact explaining **only** the necessary details. (**8 marks)**

Also provide the colab code link for each question in the report.